AIDS2024 e-Poster EPC082 Technical Details

Prioritizing High-Risk Sub-Saharan African Adolescent Girls and Young Women for Prevention Interventions Using a Bayesian Spatial Model

Technical Details

Contact: Steve Gutreuter
sgutreuter@cdc.gov

Table of Contents

1 Supplemental Methods

1.1 Model definition

1.1.1 Accommodating the survey designs

The PHIA surveys follow stratified, multi-stage sampling designs. Strata are typically defined by first-level subnational areas ("regions"). The PHIA surveys from Cameroon and Kenya were additionally stratified by urban versus rural population density. The primary sampling units are census enumeration areas ("clusters") and were randomly selected, within strata, with probability proportional to the numbers of households or population. Fixed numbers of households were selected within each cluster using systematic sampling with a random start. Additional details about all PHIA surveys can be found at `https://phia-data.icap.columbia.edu/`.

It is important to accommodate, to the extent possible, features of the survey design in Bayesian model-based estimation and prediction from survey data [1–4]. Inclusion of survey strata as a predictor strains parameter identification in the presence of the spatial error structures, and we therefore ignored the stratum identifiers in our models. Some of the PHIA surveys were also stratified by urban versus rural residency. We included the urbanicity indicator in our models.

Model-based inference from survey data predicts outcomes in sampled and non-sampled units by including the survey sampling probabilities or weights as predictors [1,5,6] and can outperform design-based estimation in terms of root mean-squared error [7]. However, our models are fitted to pooled data from 13 probability surveys. Rather than using the final blood sampling weights from each survey, we re-scaled those weights so that the country-specific sums of the re-scaled weights equaled the effective sample sizes for the surveys [8,9]. The effective sample size for a survey was the actual sample size divided by the design effect for estimation of HIV prevalence. Therefore those re-scaled weights have effective sample size as their common basis across all surveys.

1.1.2 Likelihoods

Let $y_{mij} \in \{0,1\}$, denote the absence/presence of HIV infection in young female $i = 1, \ldots n_y$ from area $j = 1, \ldots, n_a$ in country $m = 1, \ldots, 13$, and let $\boldsymbol{y} = (y_{mij})^\intercal$, where $\boldsymbol{a}^\intercal$ denotes vector/matrix transpose of $\boldsymbol{a}$. Let $\boldsymbol{p}$ denote $(p_{mij})^\intercal$, where the p_{mij} are the probabilities of infection for young female i in area j of country m. Let $\boldsymbol{Z}$ denote a matrix of "fixed"-effect demographic and behavioral covariates having coefficient vector $\boldsymbol{\beta}_Z$. Let $\boldsymbol{x}$ represent the latent vector $(x_{mj})^\intercal$, the elements of which are the population viral loads in area j of country m. $\boldsymbol{x}$ is observed indirectly through the proxy variable $\boldsymbol{w} = (w_{mjk})^\intercal$ where the $w_{mjk} = \log_{10}\left(\mathrm{VL}_{mjk} + 1\right), k = 1, \ldots, n_x$, and where VL_{mjk} is the viral load, measured in units of $\mathrm{copies} \cdot \mathrm{ml}^{-1}$ for individual mjk among n_x females and males of all ages in the corresponding areas. By definition, $\mathrm{VL}_{mjk} \equiv 0$ for HIV-negative individuals. The inclusion of $\boldsymbol{x}$ as a predictor of $\boldsymbol{y}$ requires a classical measurement error model [10] given by

$$\begin{aligned}
\boldsymbol{y} &\sim \mathrm{Bernoulli}\left(\boldsymbol{p}\right), && (1)\\
\mathrm{logit}(\boldsymbol{p}) &= \beta_0 \mathbf{1}_y + \beta_x \boldsymbol{x} + \boldsymbol{Z}\boldsymbol{\beta}_Z + \boldsymbol{b}_Y + \boldsymbol{v}_c + \boldsymbol{v}_e + (\epsilon_{mjk})^\intercal, && (2)\\
\boldsymbol{w} &= \boldsymbol{x} + (\epsilon_{Wmjk})^\intercal, && (3)\\
\boldsymbol{x} &= \alpha_0 \mathbf{1}_x + \boldsymbol{b}_X + (\epsilon_{Xmj})^\intercal && (4)
\end{aligned}$$

where β_0 is an intercept in the linear predictor (eq. 2) and $\mathbf{1}_y$ denotes a vector of n_y 1's. β_x is a hyperparameter representing the logit-linear slope in $\boldsymbol{x}$. The *iid* random vectors $\boldsymbol{v}_c \sim \mathrm{N}\left(0, \tau_c\right)$ and $\boldsymbol{v}_e \sim \mathrm{N}\left(0, \tau_e\right)$ represent country- and enumeration-area-(cluster) level random effects having precisions τ_c and τ_e, respectively, and the ϵ_{mjk} are individual-level $\mathrm{N}\left(0, \tau_y\right)$ random effects. The ϵ_{Wmjk} and ϵ_{Xmj} are independently and identically distributed (*iid*) Gaussian random errors having means 0 and precisions 10 and τ_X, respectively. The rather large fixed precision for ϵ_{Wmjk} forces $\boldsymbol{x}$ to approximate $\boldsymbol{w}$. Equations 2 and 3 comprise the observation process and eq. 4 is a latent process. This joint model contains a Bernoulli likelihood for $\boldsymbol{y}$ (eq. 1) and Gaussian likelihoods $\boldsymbol{w} \sim \mathrm{N}\left(\boldsymbol{x}, 10\right)$ (eq. 3) and $\boldsymbol{x} \sim \mathrm{N}\left(\alpha_0 \mathbf{1}_x + \boldsymbol{b}_X, \tau_x\right)$ (eq. 4). The coefficient α_0 is an intercept in the model for $\boldsymbol{x}$ and $\mathbf{1}_x$ is vector of $m \times j$ 1's. The $\boldsymbol{b}_Y$ and $\boldsymbol{b}_X$ are spatially smoothed area-level random-effect vectors (see section 1.1.3, below).

The latent Gaussian random field is then given by $(\beta_0, \boldsymbol{x}^\intercal, \boldsymbol{\beta}_z^\intercal, \alpha_0)^\intercal$. Our primary interest is in estimates of $\boldsymbol{p}$, β_x, $\boldsymbol{x}$ and $\boldsymbol{\beta}_Z$.

1.1.3 Spatial smoothing

Honoring Tobler's first law of geography [11] that "everything is related to everything else, but near things are more related than distant things", we modeled spatial correlation in HIV status and PVL using the area-level BYM2 model [12,13]. The BYM2 model extends the more popular BYM model [14] by enabling scaling which facilitates hyperprior specification [15].

The BYM2 area-level random error vectors $\boldsymbol{b}$ have the form

$$\boldsymbol{b} = \frac{1}{\sqrt{\tau_b}}\left(\sqrt{1-\phi}\boldsymbol{v} + \sqrt{\phi}\boldsymbol{u}_*\right)$$

where $\boldsymbol{v} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}\right)$, $\boldsymbol{u}_* \sim \mathrm{N}\left(\mathbf{0}, \mathbf{Q}_*^-\right)$, $\mathbf{I}$ is the identity matrix, τ_b is a precision parameter. Herein all Gaussian distributions are parameterized using precision, which is the reciprocal of variance. The hyperparameter $\phi \in (0,1)$ specifies the fraction of marginal standard error $1/\sqrt{\tau_b}$ explained by the scaled random effect $\boldsymbol{u}_*$, and $\boldsymbol{v}$ is an *iid* random effect, sometimes called the nugget. Note the the spatially independent nugget effect dominates as $\phi \rightarrow 1$ and the spatial component dominates as $\phi \rightarrow 0$. The matrix $\mathbf{Q}_*$ is a scaled version of the $mj \times mj$ spatial neighbor matrix $\mathbf{Q}$ having elements

$$Q_{gh} = \begin{cases} n_{\delta g} & \text{if } g = h, \\ -1 & \text{if } g \sim h, \\ 0 & \text{otherwise} \end{cases}$$

where $n_{\delta g}$ denotes the number of neighbors of area g, and $g \sim h$ denotes the condition that areas g and h are neighbors. The generalized variance of the random-effect vector $\boldsymbol{u}$ is given by

$$\sigma^2_{\mathrm{GV}}\left(\boldsymbol{u}\right) = \frac{1}{\tau}\exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\left[\mathbf{Q}^-\right]_{ii}\right)\right).$$

Then, $\boldsymbol{u}_*$ is obtained by scaling $\boldsymbol{u}$ such that $\sigma^2_{\mathrm{GV}}\left(\boldsymbol{u}\right) = 1/\tau_b$, the marginal variance of $\boldsymbol{b}$ [13].

We imposed sum-to-zero constraints on all BYM2 structures. The graph for $\mathbf{Q}$ has disconnected components, including singletons (isolated unitary areas) because the data spans multiple, sometimes disconnected countries, and some countries include islands. If $\tau\mathbf{R}$ is the precision matrix of $\boldsymbol{v}$, then $\mathbf{R}$ is scaled so that the marginal variances of each connected component containing at least two areal units are 1, and singletons are given an $\mathrm{N}\left(0,1\right)$ distribution.

1.1.4 Priors and hyperpriors

HIV infection is rare. We assigned vague independent $\mathrm{N}\left(0, 1/9\right)$ priors to β_0, the components of $\boldsymbol{\beta}_Z$ and α_0, which convey almost no information on the logit scale. The likelihood for $\boldsymbol{x}$ contains two *iid* Gaussian components, ϵ_X and $\boldsymbol{v}$. Any practical effect of ϵ_{Xj} is minimized by assigning fixed precision $\tau_X = 10$. A moderately informative hyperprior is required for β_x. Based on preliminary exploratory plots of survey domain estimates, we anticipated that the 20th and 80th percentiles of β_x might be approximately 4 and 7, respectively, so we chose for it a $\mathrm{N}\left(5, 1/4\right)$ hyperprior. We chose a vague Gamma(1, 0.01) hyperprior for τ_W, which gives 0.025 and 0.95 percentiles of 0.056 and 4.026 for the standard deviation [6]. Finally, we chose penalized complexity (PC) hyperpriors [12] for the BYM2 structures $\boldsymbol{b}_Y$ and $\boldsymbol{b}_X$. The PC prior for the mixing parameter ϕ will automatically shrink towards 0 (no spatial smoothing) in the absence of evidence in the data. For both, we assigned PC priors for the mixing parameters ϕ such that $\Pr\left(\phi < 0.5\right) = 2/3$, which slightly favors simpler *iid* area-level random effects. The PC priors for the area-level precisions τ_b were chosen such that $\Pr\left(\mathrm{SD} > 0.2\right) = 0.1$.

1.2 Out-of-sample predictive performance

Scientifically principled application of the model-based predictions is justified based upon the predictive performance on newly encountered AGYW, and we base predictive performance on cross-validation. Generally, ordinary cross-validation provides reliable measures of predictive performance only for single-level models which have *iid* error structures, and adaptations are needed for models having more complicated non-*iid* error structures [16,17]. However, values of the non-*iid* random effects will always be known when predictive the probabilities of infection among AGYW. The countries and districts (areas) of residence will always be known, and therefore predictions will be conditional on the responses to the risk questions, country, district-level PLV. In that case the HIV-testing outcomes $\boldsymbol{y}$ are independent conditional on the linear predictor (eq. 2) and all hyperparameters, and therefore conventional cross-validation is a viable measure of predictive performance. We used 10-fold cross-validation [18] of the final model to estimate the predictive performance on newly observed AGYW.

1.3 Computation

R version 4.3.1 [19] was used for all computations. We approximated the joint posterior distributions of the latent random field using the INLA [20] package version INLA_2023-09-09. INLA uses computationally fast nested Laplace approximations and numerical integration. Computational speed is critical to our application because of the large number of survey observations and the high dimension of the latent field, for which Markov Chain or Hamiltonian Monte Carlo sampling would have been impractical.

1.3.1 Model variations

We fitted model variations ignoring the re-scaled weights, and also followed [6] by including the re-scaled weights in *B*-spline basis functions [21] in the linear predictor (eq. 2). We chose *B*-splines having 3 df. We fitted eight model variations excluding and including country-level iid random effects $\boldsymbol{v}_c$, cluster-level iid random effects $\boldsymbol{v}_e$ and *B*-spline basis functions of the re-scaled weights (Supplemental Table 1).

Supplemental Table 1. Joint model variations.

	iid random effects		
Model name	Country	Cluster	Weight covariate
M000	No	No	No
M001	No	No	Yes
M010	No	Yes	No
M011	No	Yes	Yes
M100	Yes	No	No
M101	Yes	No	Yes
M110	Yes	Yes	No
M111	Yes	Yes	Yes

References

[1] Little RJ. To model or not to model? Competing modes of inference for finite population sampling. Journal of the American Statistical Association. 2004;99(466):546-56. doi:10.1198/016214504000000467.

[2] Gelman A. Struggles with survey weighting and regression modeling. Statistical Science. 2007;22:153-64. doi:10.1214/088342306000000691.

[3] Wakefield J, Okonek T, Pedersen J. Small area estimation for disease prevalence mapping. International Statistical Review. 2020. doi:10.1111/insr.12400.

[4] Paige J, Fuglstad GA, Riebler A, Wakefield J. Design- and model-based approaches to small-area estimation in a low- and middle-income country context: Comparisons and recommendations. Journal of Statistics and Survey Methodology. 2022;10(1):50-80. doi:10.1093/jssam/smaa011.

[5] Zheng H, Little RJA. Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. Journal of Official Statistics. 2003;19(2):99-117.

[6] Vandendijck Y, Faes C, Hens N. Prevalence and trend estimation from observational data with highly variable post-stratification weights. Annals of Applied Statistics. 2016;10(1):94-17. doi:10.1214/15-AOAS874.

[7] Chen Q, Elliott MR, Little RJA. Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. Survey Methodology. 2010;36(1):23-34.

[8] Pfeffermann D, Skinner C, Holmes D, Goldstein H, Rasbash J. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society Series B. 1998;60(1):23-40. doi:10.1111/1467-9868.00106.

[9] Rabe-Hesketh S, Skrondal A. Multilevel modeling of complex survey data. Journal of the Royal Statistical Society Series A. 2006;169(4):805-27. doi:10.1111/j.1467-985X.2006.00426.x.

[10] Muff S, Riebler A, Held L, Rue H, Saner P. Bayesian analysis of measurement error models using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2015;64(2):231-52. doi:10.1111/rssc.12069.

[11] Tobler WR. A computer movie simulating urban growth in the Detroit region. Economic Geography. 1970;46:234-40. doi:10.2307/143141.

[12] Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors. Statistical Science. 2017;32(1):1-28. doi:10.1214/16-STS576.

[13] Riebler A, Sørbye SH, Simpson D, Rue H. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Statistical Methods in Medical Research. 2016;25(4):1145-65. doi:10.1177/0962280216660421.

[14] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics. 1991;43(1):1-20.

[15] Sørbye SH, Rue H. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. Spatial Statistics. 2014;8:39-51. doi:10.1016/j.spasta.2013.06.004.

[16] Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913-29. doi:10.1111/ecog.02881.

[17] Liu Z, Rue H. Leave-group-out cross-validation for latent Gaussian models; 2023. Available from: `https://arxiv.org/pdf/2210.04482`.

[18] Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis. 2009;53(11):3735-45. doi:10.1016/j.csda.2009.04.009.

[19] R Core Team. , editor. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: `http://www.R-project.org/`.

[20] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B. 2009;71(2):319-92. doi:DOI: 10.1111/j.1467-9868.2008.00700.x.

[21] Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. Statistical Science. 1996;11(2):89-121. doi:10.1214/ss/1038425655.